CS610 Assignment 5
Name: Manish (190477)

Solution1:

S No.	Kernel Properties	Speed UP
1	1D Blocks(SIZE1)	13.2x
2	2D Blocks(SIZE2)	40.8x
3	3D Blocks(SIZE2)	49.2x

I have written 3 kernel in this problem. In first kernel Our block is 1D and is for SIZE1. We didn't get complete benefit as we can also parallelise i and k loop. In 2nd Loop I use 2D block and in 3rd kernel I use 3D blocks. I got best benefit from 3D blocks as shown in the above table. I also tried 2D blocking with loop unrolling(which is commented in the code) but my performance is not getting any better so i didn't include it into final version. All three kernel may not run in a single run on a gpu as it will result in OUT OF MEMORY ERROR. SO If you get any error of OUT OF MEMORY please comment other kernel to see the performance of that kernel. I didn't see any significant performance difference in size 8196 and 8200 for the same kernel implementation.

Solution2:

S No.	Matrix Size	Block Size	Optimisation	Speed Up(kernel1)	Speed Up(kernel2)
1	1024	8	-	268x	306x
2	1024	16	-	270x	361x
3	1024	32	-	**284x**	**350x**
4	2048	8	-	568x	658x
5	2048	16	-	569x	739x
6	2048	32	-	**595x**	**751x**
7	4096	8	-	682x	794x
8	4096	16	-	679x	**920x**
9	4096	32	-	**693x**	893x
10	1024	32	4-Way-Loop Unrolling	**314x**	**503x**
11	2048	32	4-Way-Loop Unrolling	**573x**	**953x**
12	4096	32	4-Way-Loop Unrolling	**717x**	**1246x**

We can not that in most of the cases Block size =32 is giving us the best performance. If w e apply 4 way loop unrolling then our performance is further enhanced for almost all the cases.

Solution3:

Comparing the 4 different Kernel Performance:

S No.	Type	Time Taken By Kernel	Time taken by CudaMemCpy
1	CudaMalloc	141ms	143ms
2	CudaHostAlloc(cudaHostAllocDefault)	157ms	157ms
3	CudaHostAlloc(cudaHostAllocMapped)	160ms	159ms
4	CudaMallocManaged()	165ms	164ms

I ran the experiments mutiple times. Results for CudaHostAlloc(cudaHostAllocDefault), CudaHostAlloc(cudaHostAllocMapped) and CudaMallocManaged() are not consistent. But what I observe is that CudaMalloc always run faster than the rest of the three. The results are not consistent because at the time of experiment heavy load is also running on the gpu2 which may alter the results. Nvprof only work on gpu2. The load is running for a long time so I am not able to capture when gpu is free completely.
CudaMalloc NVPROF OUTPUT

final1.txt

```
==864264== NVPROF is profiling process 864264, command: ./binary
==864264== Profiling application: ./binary
==864264== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   97.18%  136.30ms         1  136.30ms  136.30ms  136.30ms  kernel2(unsigned long const *, unsigned long const *, unsigned long
                    2.15%  3.0212ms         2  1.5106ms  1.4204ms  1.6008ms  [CUDA memcpy HtoD]
                    0.67%  941.25us         1  941.25us  941.25us  941.25us  [CUDA memcpy DtoH]
      API calls:   65.25%  273.61ms         3  91.204ms  84.136us  273.44ms  cudaMalloc
                   34.26%  143.66ms         3  47.886ms  1.5816ms  138.48ms  cudaMemcpy
                    0.26%  1.0814ms       404  2.6760us     104ns  195.59us  cuDeviceGetAttribute
                    0.18%  767.27us         3  255.76us  195.68us  289.58us  cudaFree
                    0.02%  86.431us         4  21.607us  16.359us  36.419us  cuDeviceGetName
                    0.01%  40.685us         1  40.685us  40.685us  40.685us  cudaLaunchKernel
                    0.01%  25.187us         2  12.593us  8.4690us  16.718us  cudaEventRecord
                    0.00%  14.294us         1  14.294us  14.294us  14.294us  cudaEventElapsedTime
                    0.00%  13.351us         4  3.3370us  1.7350us  7.0720us  cuDeviceGetPCIBusId
                    0.00%  10.376us         2  5.1880us     643ns  9.7330us  cudaEventCreate
                    0.00%  4.4220us         1  4.4220us  4.4220us  4.4220us  cudaEventSynchronize
                    0.00%  2.2070us         8     275ns     114ns     731ns  cuDeviceGet
                    0.00%  1.7590us         3     586ns     203ns  1.3210us  cuDeviceGetCount
                    0.00%  1.1360us         4     284ns     219ns     470ns  cuDeviceTotalMem
                    0.00%     725ns         4     181ns     138ns     238ns  cuDeviceGetUuid

```

CudaHostAlloc(cudaHostAllocDefault) NVPROF OUTPUT

final2.txt

```
==864430== NVPROF is profiling process 864430, command: ./binary
==864430== Profiling application: ./binary
==864430== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   51.20%  157.90ms         3  52.634ms  1.2202ms  154.17ms  [CUDA memcpy HtoH]
                   48.80%  150.47ms         1  150.47ms  150.47ms  150.47ms  kernel2(unsigned long const *, unsigned long const *, unsigned l
      API calls:   66.82%  321.43ms         3  107.14ms  2.5763ms  316.26ms  cudaHostAlloc
                   32.83%  157.95ms         3  52.649ms  1.2287ms  154.19ms  cudaMemcpy
                    0.23%  1.1079ms       404  2.7420us     127ns  202.64us  cuDeviceGetAttribute
                    0.08%  361.03us         1  361.03us  361.03us  361.03us  cudaEventSynchronize
                    0.02%  81.433us         4  20.358us  16.538us  31.013us  cuDeviceGetName
                    0.01%  36.065us         2  18.032us  15.454us  20.611us  cudaEventRecord
                    0.01%  34.082us         1  34.082us  34.082us  34.082us  cudaLaunchKernel
                    0.00%  22.515us         2  11.257us     815ns  21.700us  cudaEventCreate
                    0.00%  13.013us         4  3.2530us  1.7060us  6.9400us  cuDeviceGetPCIBusId
                    0.00%  11.272us         1  11.272us  11.272us  11.272us  cudaEventElapsedTime
                    0.00%  4.8270us         3  1.6090us     693ns  3.4390us  cudaFree
                    0.00%  1.8180us         3     606ns     177ns  1.4250us  cuDeviceGetCount
                    0.00%  1.7880us         4     447ns     250ns     911ns  cuDeviceTotalMem
                    0.00%  1.4690us         8     183ns     126ns     524ns  cuDeviceGet
                    0.00%     891ns         4     222ns     161ns     359ns  cuDeviceGetUuid

```

CudaHostAlloc(cudaHostAllocMapped) NVPROF OUTPUT

final3.txt

```
==864559== NVPROF is profiling process 864559, command: ./binary
==864559== Profiling application: ./binary
==864559== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   50.75%  159.87ms         3  53.290ms  1.4262ms  156.72ms  [CUDA memcpy HtoH]
                   49.25%  155.15ms         1  155.15ms  155.15ms  155.15ms  kernel2(unsigned long const *, unsigned long const *, unsigned l
      API calls:   67.86%  342.57ms         3  114.19ms  2.6510ms  337.18ms  cudaHostAlloc
                   31.68%  159.93ms         3  53.309ms  1.4395ms  156.73ms  cudaMemcpy
                    0.25%  1.2377ms       404  3.0630us     105ns  221.95us  cuDeviceGetAttribute
                    0.16%  822.01us         1  822.01us  822.01us  822.01us  cudaEventSynchronize
                    0.03%  128.03us         4  32.006us  18.085us  67.584us  cuDeviceGetName
                    0.01%  38.158us         1  38.158us  38.158us  38.158us  cudaLaunchKernel
                    0.01%  33.930us         2  16.965us  13.755us  20.175us  cudaEventRecord
                    0.01%  26.536us         2  13.268us     871ns  25.665us  cudaEventCreate
                    0.00%  19.972us         4  4.9930us  1.5790us  14.765us  cuDeviceGetPCIBusId
                    0.00%  9.3950us         1  9.3950us  9.3950us  9.3950us  cudaEventElapsedTime
                    0.00%  4.6850us         3  1.5610us     560ns  3.4880us  cudaFree
                    0.00%  3.4440us         8     430ns     161ns  1.8380us  cuDeviceGet
                    0.00%  3.1930us         3  1.0640us     190ns  2.2610us  cuDeviceGetCount
                    0.00%  2.1000us         4     525ns     351ns     782ns  cuDeviceTotalMem
                    0.00%  1.3140us         4     328ns     248ns     432ns  cuDeviceGetUuid

```

CudaMallocManaged() NVPROF OUTPUT

final4.txt

```
==864725== NVPROF is profiling process 864725, command: ./binary
==864725== Profiling application: ./binary
==864725== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   83.61%  130.63ms         1  130.63ms  130.63ms  130.63ms  kernel2(unsigned long const *, unsigned long const *, unsigned long*)
                   11.59%  18.102ms         2  9.0509ms  9.0369ms  9.0649ms  [CUDA memcpy HtoD]
                    4.81%  7.5123ms         1  7.5123ms  7.5123ms  7.5123ms  [CUDA memcpy DtoH]
      API calls:   64.07%  301.77ms         3  100.59ms  10.828us  301.70ms  cudaMallocManaged
                   35.02%  164.94ms         3  54.981ms  11.408ms  140.71ms  cudaMemcpy
                    0.42%  1.9723ms         1  1.9723ms  1.9723ms  1.9723ms  cudaEventSynchronize
                    0.24%  1.1480ms       404  2.8410us     113ns  221.49us  cuDeviceGetAttribute
                    0.20%  924.01us         3  308.00us  286.34us  339.08us  cudaFree
                    0.02%  87.627us         4  21.906us  16.248us  37.420us  cuDeviceGetName
                    0.01%  30.416us         2  15.208us  13.943us  16.473us  cudaEventRecord
                    0.01%  28.570us         1  28.570us  28.570us  28.570us  cudaLaunchKernel
                    0.00%  19.114us         2  9.5570us     671ns  18.443us  cudaEventCreate
                    0.00%  14.815us         4  3.7030us  1.8520us  8.4570us  cuDeviceGetPCIBusId
                    0.00%  11.361us         1  11.361us  11.361us  11.361us  cudaEventElapsedTime
                    0.00%  10.079us         4  2.5190us     174ns  9.3330us  cuDeviceGetUuid
                    0.00%  2.3740us         8     296ns     118ns     516ns  cuDeviceGet
                    0.00%  1.9350us         3     645ns     167ns  1.4040us  cuDeviceGetCount
                    0.00%  1.3850us         4     346ns     229ns     676ns  cuDeviceTotalMem

==864725== Unified Memory profiling result:
Device "NVIDIA GeForce GTX 1080 Ti (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
      53         -         -         -           -  5.241901ms  Gpu page fault groups

```

To run the experiments please uncomment the corresponding lines in the file and run for that version. For CudaMalloc[127-129 line], CudaHostAlloc(cudaHostAllocDefault)[132-134 line], CudaHostAlloc(cudaHostAllocMapped)[136-138 line],CudaMallocManaged()[140-142 line]

Solution4:

S No.	Valur Of N	Speed Up(kernel1)	Speed Up(kernel2)
1	64	2.1x	3.3x
2	128	2.4x	4.1x
3	256	2.0x	2.9x
4	512	2.24 x	2.11x

The Optimised version exploits spacial locality by tiling and it gives improved and consistent performance for all N.
The optimised version exploits spacial locality but for boundary points, it go for the global memory and this is one of the possible bottleneck because of which we didn't get much improvement in this scenario.

Run.sh Contain all compilation commands for all questions